DØ Grid Production Computing Initiative
Closure Report

V1.2.1
11 July 2008
Robert D. Kennedy, *et al.*

Table of Contents

Initiative Overview

Introduction

The DØ Grid Production Computing Initiative is an umbrella project to achieve a broad set of goals by applying a modest level of project management formalism to track and sustain progress. The Initiative adds staff to help plan, prioritize, and coordinate work amongst the existing personnel and groups, with an eye towards identifying and reducing the long-term maintenance and support requirements of RUN2 experiments.

CD Charge to the Initiative

The charge to the Initiative from the Computing Division can be summarized by "do what is important in six months to get DØ Grid Computing up to production grade". Based on consultation with the major stakeholders, existing work plans, and consideration of the DØ SAM-Grid Prioritization (26 March 2007) and the DØ "Charge to SAM-Grid-DØ Project Manager" (31 April 2007) documents[1], a number of objectives were identified to support this charge, listed below. The Initiative Objectives are:

1. DØ Grid Production Architecture
2. Adjust Responsibilities among SAM-Grid, DØrunjob, DØrepro, automc tools
3. Transition Operations Support from SAM-Grid Dev to Run2 Ops
4. DØ Primary Production on Fermigrid
5. Extend Grid Production System Functionality (if necessary)
6. Preparation of DØ Apps for Long-term
7. Evaluation of DØ SAM-Grid Requests

More realistic estimates of task durations and staff availability from early Initiative experience indicated that not all of the envisioned work could be fit into the Initiative. Major tasks were prioritized, and lower priority work was dropped or reduced in scope.

Project Team

The Initiative was led by Amber Boehnlein (CD ADS Dept Head, DØ) in the Project Director role and supported by Robert D. Kennedy (CD OPMQA) in the Project Manager role. The Coordination Committee consisted of Eileen Berman (CD Grid Facilities Dept Head), Bill Boroski (CD Associate Head for PM&QA), Mike Diesburg (DØ Production Coordinator), Gabriele Garzoglio (CD OSG Group Leader, SAM-Grid Development Leader), Qizhong Li (CD REX Dept Deputy Head), Adam Lyon (CD Rex Ops Group Leader, SAM Project Leader, DØ). Also contributing to the Initiative were the DØ collaborators Peter Love (dØrunjob maintainer), Joel Snow (DØ MC Production Coordinator and automc maintainer), and Daniel Wicke (dØrepro-tools maintainer). The DØ Offline Computing liaison role was performed by Amber Boehnlein, who became the co-leader of DØ Offline Computing during this Initiative.

[1] Both documents as well as a supporting diagram are available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/d0-grid-production-computing-initiative/background-docs

Project Repository

Project Management and some subject matter documentation are maintain at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative . This area is organized into the following folders:

- Coordination Meetings – Slides, minutes, and other artifacts from the Initiative Coordination Meeting held biweekly, then weekly.
- DØ Primary Production on Fermigrid Meeting – Slides, pictures, notes, and other artifacts from the DØ Primary Production on Fermigrid Meeting held 24 August 2007, 0900-1145.
- Initiative Baseline Planning Meeting – Meeting 0900-1300 on 04 October 2007 to review candidate baseline plan v0.9.9. Slides and project plan artifacts.
- Initiative Wind-down Summaries – Slides and other documents summarizing the DØGPCI as it nears and reaches its conclusion from close-out related meetings in early to mid-February 2008.
- Initiative Close-Out Documents – Lessons Learned Meeting, Closure Report, and related documentation from June 2008.
- Project Management Docs – Documents describing the project management aspects of the DØ Grid Production Computing Initiative. This includes MS Project schedules and related reports updated frequently throughout the Initiative.
- Subject Matter Docs – Documents tracked by the Initiative that describe the DØ Grid Production System and DØ Applications. Many of these are the high-level subject matter documentation deliverables of the Initiative.
- Background Docs – Documents and sources describing the DØ Grid Production System, DØ Applications, Customer Requirements and Priorities. This is not meant to replace DØ experiment repository, however, only to make available some documents related to the Initiative in a more public repository.

Project History

Motivation and Development: March 2007 – April 2007

- March 2007: DØ SAM-Grid Development Prioritization[2] (26 March 2007) – from DØ Collaboration
- April 2007: DØ Grid Production Workshop (11 April 2007) – URL: http://cd-docdb.fnal.gov/cgi-bin/DisplayMeeting?conferenceid=335
- Charge to SAM-Grid-DØ Project Manager[3] (31 April 2007) – from DØ CPB
- May 2007: Initiative concept developed: Umbrella Project to last 6 months.

Planning: June 2007 – October 2007

- June 2007: CD FY08 Budgeting slows planning. Existing plans still executing.
- July 2007 Coordination meetings begin. Execution tracked in parallel with formal planning. “Road to Operations” by Gabriele Garzoglio and “Automation of

[2] https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/d0-grid-production-computing-initiative/background-docs/samgridprioritization_041307.doc

[3] https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/d0-grid-production-computing-initiative/background-docs/amberchargev3.doc

Health Alarms" by Andrew Baranovski integrated into the Initiative schedule.
- August 2007: DØ Primary Production on FermiGrid Planning Meeting – identified need for a supported system diagram and service-hardware mapping.
- September 2007: Start Ops responsibility transfer in phases. 4 Major Milestones already achieved due to early successes and pre-Initiative work in progress.
- October 2007: Baseline Planning meeting and v1.0.0 project plan release.

Execution: July 2007 – February 2008

- Planned to execute from July 2007 to Jan 2008, then extended to Feb 15, 2008.
- Biweekly, then weekly in early 2008, status gathering and coordination meetings.
- Execution effort limited early by Operations support load, though this was reduced over time as operations reduction tasks began to pay off.
- Nov/Dec 2007: Work slow-down against plan. Some major tasks slip past Feb 15.
- February 2008: Overall 2X reduction in operations issues reported per unit time, 3-4X reduction in major issues reported, comparing email reports in October 2007 to those in late January/early February 2008. Signoff by DØ production and MC production coordinators that grid production stability achieved.
- 15 February 2008: End Initiative formally, but track the remaining open tasks to completion and consult on longer-term processes.

Close-Out: March 2008 – June 2008

- March –April 2008: Less frequent status gathering and coordination meetings. Participating groups interleave other work with Initiative per FY08 effort plans.
- April: Enough open tasks judged to be done, close-out to resume in May 2008.
- 06 May 2008: Final status call meeting.
- 03 June 2008: Final "Lessons Learned" meeting.
- 27 of 29 Major Milestones completed.
- 2 of 29 Major Milestones were "Closed Incomplete":
    - Mile 14 (DØrepro-tools): Unlikely to be achieved soon due to external resource availability. DØrepro-tools was not adapted to use SAM v7.
    - Mile 28 (Production on FermiGrid Stable): Completed, but more work is required to achieve goal of long-term stability at agreed service levels.

Reasons for Closing the Project

The DØ Grid Production Computing Initiative was defined to be a time-limited project to achieve what was possible in six months, effectively beginning in July 2007. In early February 2008, after an approved modest extension, the Initiative had accomplished its charge as best as possible in approximately the time allotted. With the DØ production coordinators sign-off on operations improvement goals on 11 February 2008, we presented the Initiative accomplishments and a status report of the DØ SAMGrid Development Requests at an executive meeting on 13 February 2008 to the DØ spokespeople and Computing Division Head. Included was a well-received proposal to track all remaining open tasks to completion, since the remaining timeline would stretch out as participating groups were committed to devote effort to other projects as well. All open major tasks were completed by 06 May 2008.

State of the Initiative Objectives

1. DØ Grid Production Architecture

Goals:

- Document the high-level architecture of the DØ Grid Production system in order to clarify roles and responsibilities among the services and tools. The nature of this document is defined by Vicky White, who will sign off on the deliverable.

State at Closing: DONE.

Deliverables:

- High-Level Architecture Document to serve this objective, at https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/arch_high_level_v2_draft.pdf
- Support document with more DØ-specific detail, at https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/D0-gridproductionarchitecture-v1-2.pdf

Outstanding Risks (including Operations, Support): None identified.

2. Adjust Responsibilities among SAM-Grid, DØrunjob, DØrepro, automc tools

Goals:

- Streamline these tools to insure that each was an independent of the others as much as was reasonably possible at this point in their life cycle.
- Define procedures for component, integration, and system testing to minimize downtime due to “testing in production”.
- Accomplish some related DØ-requested development requests. The work was selected based on informal cost-benefit evaluation.

State at Closing: MOSTLY DONE.

The adaptation of DØrepro-tools to SAM v7 request system was worked on by Daniel Wicke, but was not completed. A defect in a SAM Python interface was reported by Daniel in October 2007, which was not resolved until December 2007. By that time, Daniel has other commitments and was unable to devote time to this task.

Deliverables:

- Rework of the SAMGrid-DØRunjob interfaces and responsibilities to make each tool more independent of the other. Example: decoupling of DØrunjob macro and SAMGrid. The benefits were judged to not be worth the effort required to accomplish a complete decoupling of the tools at this point in their life cycle.
- Procedures for integration testing to improve component reliability before

integration begins, and system reliability before production deployment.
- Establishment of a SAMGrid test stand to enable as much feature testing as possible of new code versions before production deployment.
- Feature: Ability to start production from Stage 2, and error recovery in this kind of phased processing. It was agreed in July 2007 by Initiative, CD, and DØ representatives that being able to start production at later stages was not worth the effort required, especially as this would be eventually simplified by development to support generalized job types.
- Feature: Support for generalized job types based on I/O characteristics.
- A document describing how to add production paths in the future (a major DØ request), based on the new support for generalized job types, available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/futureproductionpaths.doc . The steps required are listed in a generic WBS template for the work required to adapt a DØ application to Grid Production, available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/adaptD0app-genericwbs-v1_4.doc

Outstanding Risks (including Operations, Support):
- The new support for generalized job types has not been thoroughly tested by actually adapting a DØ application to run in grid production. The CD/SCF/GRID/OSG group recognizes there may be additional consulting and development work required when this feature is finally tested in the field.
- There is some risk of additional support required at a later date due to DØrepro-tools not yet using the SAM v7 interface.

3. Transition Operations Support from SAM-Grid Dev to Run2 Ops

Goals:
- Document aspects of the Grid Production system as requested by DØ.
- Implement the tools, create procedures, perform training, and write documentation to transfer operations support from the expert SAM-Grid development team (the CD/SCF/GRID/OSG Group) to the Run2 operations team (the CD/SCF/REX/OPS group).
- As a prerequisite to operations transfer, reduce the effort required to operate the grid production system. Automate as many monitoring tasks as possible.
- Perform a systematic re-installation of Grid Production services on performant new hardware in a standard and documented configuration.
- Define sustainable support processes and communications channels.
- Apply “production” procedures and mindset to the operation and management of the DØ Grid Production System. Prepare disaster recovery procedures.

State at Closing: DONE.

Deliverables: A partial list of what was accomplished –

- Forwarding Node Installation documentation, integrated into the SAM-Grid installation manual, is available at: https://plone3.fnal.gov/SAMGrid/Wiki/ForwardNodeInstallationInstructions .
- Grid Operations Policies document: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/samgridops_v1_2_final.pdf
- Following a detailed plan, primary operations support was transferred to the REX/Ops group successfully. Developers play a consultant back-up role in the documented support process. Operators and developers meet regularly to review open requests and augment existing procedures where gaps are identified.
- All support requests are tracked in SAM-IT, a plone-based issue tracking tool.
- 3 of 3 aging Forwarding Nodes and 1 of 1 aging Queuing Node were successfully replaced with new performant hardware, and installed in newly-defined standard configuration with upgraded infrastructure software (VDT).
- Installation of LCG Forwarding Nodes in a standard configuration was pursued, and was about to be accomplished on at least one site in the UK at this writing.
- A basic SAM-Grid Test Stand was created.
- Routine maintenance operations were automated. Examples include job queue clean-up and disk space (logs, output sandboxes) clean-up.
- The average output sandbox size was reduced without loss of usability.
- System health alarms set up for critical services in the Grid production system.
- A basic system testing infrastructure was created to allow simple emulations of user jobs to be run periodically as an end-to-end test of critical services.

Outstanding Risks (including Operations, Support):

- The new processes for handling support requests need to be followed diligently in order to become ingrained habit, “how we do things”. If shortcuts and exceptions are taken for non-critical/emergency requests, then corrective action or intervention may be required in order to help reinforce these processes.
- Hardware upgrades may be required to be done at least one more time before the end of Run2 data analysis. Software upgrades will surely have to be done several more times. The effort to automate, simplify, and document operations should continue. This objective is the start of a long-term process, not the end of one. If this is not recognized, then operations will become ever more effort-intensive for operators and the system ever less reliable for users over time.
- Some issues will arise that will still require substantial effort from developers, such as the 32k sub-directory limit on ext2/3 file systems and its impact on grid tools at greater usage levels. These will still need to be addressed, with a process to fill the longer timescale tracking role played by the Initiative.
- The SAM-Grid Test Stand will need to be maintained and extended. The process to use it to validate versions before production release depends on its viability.
- The system health alarms and monitoring will require maintenance over time as the grid production system components change. Infrequent scheduled inspections may help catch inconsistencies. Once the monitoring system becomes out-of-sync or is left unused for a time, the benefits to operations will decrease substantially.

4. DØ Primary Production on Fermigrid

Goals:

- Evaluate and address the needs of DØ Primary Production in order to establish stable, predictable service.
- Support the transfer of DØ Primary Production to FermiGrid.
- Simplify the task of running production processing as much is possible.

State at Closing: MOSTLY DONE.

While this work was signed off by DØ production coordinators as successful at the end of the Initiative on 11 February 2008 (https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/initiative-close-out-and-summaries/D0gpci-20080211-D0ppf.ppt), demand on the system increased shortly afterward and operations issues arose which severely impacted operational efficiency. This made clear the need for a service level agreement and/or an automatic throttle to limit demand to pre-determined stable usage limits in order to maintain stable production operations over the long term. Neither of these were planned deliverables of the Initiative, however. We now believe that having at least one of these in place is necessary to accomplish this goal, so we have chosen to declare this objective to not be fully achieved, to encourage future work in this area.

Deliverables: Most work identified to support this objective was tracked under other objectives. A few distinct deliverables include:

- Low-Level Workflow/Dataflow Diagrams: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/D0-grid-production-system-workflow-dataflow/
- Hardware-Service Mapping: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/D0gpci-services-and-nodes-3.doc
- Feature: Forward user's job scheduler requirements to sites, to allow user control similar to functionality of ReSS, but not yet offered by ReSS in detail.

Outstanding Risks (including Operations, Support):

- The hardware-service mapping should be reviewed from time-to-time since the hardware used in various roles changes with limited notice. This mapping proved to be a very useful tool for supporting effective communication between diverse groups working together, but will remain useful only if it is kept accurate.
- After much was done to improve the capacity and robustness, the grid production system was shortly thereafter suffering a high rate of operations problems at the same time that DØ production was achieving record production levels. A Service Level Agreement and/or automated demand throttle is needed to balance the demand on the grid production system with its capabilities, in order to insure predictable service with reasonable operations costs.

5. Extend Grid Production System Functionality

Goals:

- Evaluate requests to extend the existing Grid Production system functionality. If a requested extension is judged worthwhile based on cost-benefit analysis and possible to complete in six months, then implement and deploy the extension.

State at Closing: DONE.

Deliverables: No extensions were deemed both necessary and feasible to accomplish.

- Two functionality extensions were evaluated during the Initiative: "Forwarding Nodes behind Firewalls" and "Minimal Resource Brokering". Documentation is available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/evaluation-of-the-extension-to-development-v1_0.doc .
- A more thorough summary of SAM-Grid Brokering as it relates to the Minimal Resource Brokering topic and DØ SAM-Grid Development request (production #9) is documented at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/executivesummaryofsam-gridbrokeringrequest.doc .

Outstanding Risks (including Operations, Support): None identified.

6. Preparation of DØ Apps for Long-term

Goals:

- Propose best practices to help "future-proof" critical DØ applications (and data) against hardware failures, personnel transitions, and other environmental factors.

This was originally envisioned with a broader goal to coordinate implementation of best practices for critical DØ applications, or at least provide a detailed evaluation of what is required for that implementation. This is consistent with the theme of CD Run2 Initiatives, to help prepare Run2 Computing for reduced collaboration effort in computing support over the next few years while maintaining the same service levels. We found the original goal to be quite ambitious given the DØ environment. Some heavily invested maintainers did not see the need to prepare now for shifter-oriented operations, risk reduction, and standardized practices. Not all D0 applications lend themselves to pure shifter operation, limiting the benefit of their preparation. Some critical D0 applications were not even maintained in CVS. We concluded that the implementation of best practices was likely to take much longer than the Initiative lifespan and was less likely to deliver immediate value compared to other work in the Initiative. We scaled back this goal to be less intrusive on DØ application maintainers, but still having value to both CD and DØ managers: a best practices checklist for critical DØ applications.

State at Closing: DONE. (Judged against the revised goal)

Deliverables:

- Recommendations for the preparation of DØ Applications for reduced long-term maintenance are documented in: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/genericapprecommendations-v1_0.doc

- List of known portability issues for DØ Applications, with respect to adaptation to the Grid Production System: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/d0-grid-production-computing-initiative/subject-docs/adaptd0app-portabilityissues-v1_0.doc

Outstanding Risks (including Operations, Support):

- There is a potential for significantly more operations requests from the experiment to the Computing Division as support levels for computing from the collaboration drop due to Run2 operations ending and LHC operations beginning.
- There is a potential for a critical DØ application being left in unmaintained if its maintainer leaves without performing a thorough application transition process.

7. Evaluation of DØ SAM-Grid Requests

Goals:

- Address the topics listed in the "DØ SAM-Grid Development Prioritization" document (which lists the 17 DØ SAM-Grid Requests) that were appropriate for the Initiative and were high enough priority relative to other work undertaken.
- Evaluate at some level all of the topics listed in the "DØ SAM-Grid Development Prioritization" document to establish a foundation for future work.

State at Closing: DONE.

Deliverables:

- Evaluation of the "DØ SAM-Grid Development Prioritization" requests and their state at the end of the Initiative is documented in: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/D0gpci-status_D0_requests-v1_0.doc. The Initiative was not intended to undertake work to accomplish all of these requests, only to evaluate each, working those which were feasible within the Initiative constraints.
- Created a generic WBS template for the work required to adapt a DØ application to Grid Production, available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/adaptD0app-genericwbs-v1_4.doc
- Applied the generic WBS to the adaptation of Stand-Alone Recocert to Grid Produciton, available at: https://plone4.fnal.gov/P0/CD-OPMQA/project-management-activities/D0-grid-production-computing-initiative/subject-docs/adaptD0app-sarecocertwbs-v1_4.doc

Outstanding Risks (including Operations, Support):

- Given the historical sequence of events, some DØ collaborators may have believed that the "DØ SAM-Grid Development Prioritization" document described work that was to be undertaken by the Initiative. We believe there was sufficient communication of Initiative goals to minimize such misunderstandings.

Lessons Learned

Grid Technology

Monitoring

- The monitoring from the point of view of users still needs much work to enable users to be as effective as they were with traditional batch systems, and reduce the rate of “little” questions to operators. This leads to increased demand on the operations team and reduced confidence in the system for users. This is partly an issue for the underlying grid tools, but not entirely. The grid system is composed of different tools which report state differently. A system-wide monitoring layer that presents a single unifying picture to the user might pay for itself with a reduction in “little question” investigation and responses by operators.
- Deploying improvements to monitoring earlier in the Initiative would have not only enhanced the usability for the system sooner, but also helped put development and operations priority decisions on firmer ground.

Grid Use by Global Collaborators

- DØ Production coordinators pointed out a gap in the Grid Computing from the point of view of large customers: there are no tools yet to automate resource brokering across multiple grids. Global collaborations like DØ need to be able to coordinate the use of resources available on different grids to accomplish very demanding and critical production tasks. DØ MC Grid Production coordinators have to manually manage their job submission across a combination of OSG and LCG sites, which is labor intensive and leads to resource utilization inefficiencies. This is a hard problem in general, and beyond the Initiative resources to address.

Distributed Systems Management:

- Coordination with external resources/people was sometimes difficult and unreliable. Motivation for experimenters, especially remote researchers, comes more from the attainment of new features with short-term benefits to physics analysis. This contrasts with the goal of RUN2 Initiatives to reduce long-term support requirements.
- Maintaining a standard configuration for critical services at remote sites is awkward unless there is cooperation from the leadership at the remote site. Any one site can insist its individual configuration be support unless proven otherwise, challenging attaining a standard configuration for production operations. Working more closely with the experiment management to interface with such sites will reduce the impact of such sites on Initiative personnel time and enable a wider variety of responses.
- Remote resources/people involved in work related to the Initiative were eager to contribute positively to the DØ Grid Production system. Where there were disagreements, the issues were priorities, cost/benefit evaluations of tasks, a mutually agreeable schedule, or communications, not whether or not to help. Integrating Initiative communications and decision-making closer to the established experiment processes that remote contributors are already engaged in

may have helped reduce misunderstandings in these areas.

General Management:

- Grid Technology, as with any new technically complex technology, presents a management challenge. There is a “specialize versus cross-train” tension where operations teams have an uneven level of experience with Grid middleware.

Production Operations Support

General:

- The best benefit/cost task is generally agreed to have been the health monitoring and alarms. It would have been advantageous to implement these as soon as possible, which is consistent with the point made elsewhere of improving monitoring as early as possible to leverage as much as possible in other tasks.
- Starting the ticket meetings sooner in this Initiative might have helped the support transfer process proceed more quickly. Perhaps such meetings could have started as early as the handoff of the first operations checklist. This point is not universally agreed. The operations team was heavily burdened at this time and perhaps could not have devoted enough consistent attention to make efficient use of the meetings at that time.
- Defining what is meant by Production Operations support, for both operators and developers, is crucial and requires follow-up to avoid shortcuts to meet schedule.
- It is suggested that we remeasure the tickets-per-month level (as was done in February 2008) and quantify the load reduction on development and operations staff. Goals are to determine if the system if performing well and quantify how much the Initiative has reduced the operations load. Caveats: demand level changes and is not well-measured by a single number yet, personnel have moved to lower priority work thus never appearing to become free.
- Tickets should be assigned to a few gross severity categories to distinguish five minute simple tasks from major long-term design/platform issues. Severity should reflect a combination of the potential impact on the system and/or user’s work, estimated effort to resolve, and the level of understanding of the underlying issue.

Service Level Agreement:

- A Service Level Agreement backed up by monitoring of the criteria used in the agreement, is crucial for the delivery of a successful production system. As occurred in this Initiative, performing substantial and complex upgrades to meet verbally stated demands can end up being judged something of a failure once user demand increases to make use of apparently increased system capacity. The Initiative should have pursued at least a general SLA very early on with the customer, made sure it could be monitored, before it took on a substantial hardware and software system upgrade to better serve the customer. This was discussed as early as September 2007, but it was considered a lower priority at the time compared to retrofitting existing hardware and support transition tasks.
- Despite being reinforced by Initiative experience, it is still not clear what limit dimensions or values might be used since, until a system starts to fail, it is not

very predictable where or what its limits are. Nevertheless, selecting a few criteria to monitor with soft statements about future capacity is far better than no agreement at all. Getting the process started and expectations set is the first priority. Fine-tuning the quantitative aspects can follow in an iterative process.

- With an SLA in place and monitoring to track it, one can start to set expectations with the customer that there are real costs to increased capacity with a semi-quantitative discussion. If a 50% increase in capacity over agreement is requested, then (example numbers only) it will cost the customer 2 additional forwarding nodes, 1 additional server-class node to break out some central services to spread load, 6-8 FTE-weeks of operator time to procure, prepare, and install the new nodes, and 1 FTE-week of developer time to consult on an optimal configuration. Further, there is at most 6 month lag time between a funded formal request and its implementation. Otherwise, the existing service levels remain in effect.

Large Customer, Distributed Operations:

- For optimally effective production operations, there should be clearly defined responsibilities between CD and DØ for all of the software and hardware systems involved in the distributed Grid Production system, both on-site and remote. It was unclear what the institutional commitments are for software products and services at DØ, especially for remote services, some of which are critical for day-to-day operations. Many problem reports involve remote systems where the response coordination and time is severely impacted by the time it takes to determine who is willing to administer the remote service (if anyone) and achieving agreement with that remote service administrator on related issues such as a common, maintainable configuration that have already been worked out with the remainder of the DØ Grid Production system administrators. This situation will get worse as individuals who have graciously volunteered their time to work on remote systems in the past will eventually move on in their career and no longer be able to help. One proposal: All services are be backed by institutional MOUs, otherwise a service is removed from the "production" system.

Initiative Processes

Tracking:

- Rob's tracking of the details was good, but reporting at the Coordination Meetings was perhaps too detailed. The combination of e-mail status requests and face-to-face discussion to cover more complex topics seemed to work.
- The Coordination Meetings seemed to serve both high-level topics (priorities, decisions) as well as low-level task coordination. The e-mail status reports were often thin, requiring follow-up. Perhaps moving some of the detailed follow-up to the end of the meeting would help address this point, if adequate time is reserved.
- It was not clearly determined which MS Project reports had value for the participants or for the stakeholders. Reducing the effort to produce unused reports could have freed more time for communication and coordination activities.
- While the winter holidays vacation effect on task durations was considered ahead of time, the increasing uncertainty of plans beyond the initially well-thought-out few months was not. The Initiative began to stall in November 2007-January 2008

as the original detailed plans were accomplished and the next layer of work began. Duration uncertainties increased, and priorities became less clear. With less cohesion between tasks, and other work vying for participants' attention, the fraction of effort devoted to Initiative tasks seemed to waiver as well.

Communications:

- Communications improved with more DØ collaborators and CD-DØ representatives at the Coordination Meetings, representing the breadth of stakeholders and roles on the experiment.
- An explicit communications plan may have helped identify gaps in the communication chain, especially with groups within the DØ experiment. A balance has to be struck though in a small group like the Initiative between broad communications and maintaining focus on accomplishing tasks which will rarely be unanimously viewed as being the highest priority work to be done.
- While FermiGrid Services group representation (other than by Eileen) was not necessary at each Coordination Meeting, having a higher level one-page summary of the Initiative plans would have kept them posted on coming expectations. Perhaps a page with links for drill-down into different stakeholder "attention" lists. The same applies for FEF, and does this scale? Counter-point: would they have read a passive web page and/or e-mail in their busy communications stream? Some communications should have continued to be performed by the task leaders working with the FermiGrid Services group, and perhaps that expectation was not clear. Perhaps setting aside infrequent Coordination Meetings with topics of interest to a specific group would have addressed both points, with a specific invitation and agenda available ahead of time with some background information.

Organizational Responsibility:

- Tasks that involved significant work with a broad set of application maintainers within the DØ experiment suffered from unclear application responsibilities. It took over a month to get to the point where meaningful e-mail could be exchanged with an application person. In some cases, application maintainers either believed they only had responsibility for operating an application process or believed the application to be frozen with no motivation for change in the foreseeable future. This made working on some Initiative tasks extremely challenging and seemingly unproductive.
- Some critical tools in DØ Offline are maintained by authors on a goodwill basis without an MOU to insure support continuity. While the maintainers' sometimes volunteer efforts are very much appreciated by all involved, this presented significant short-term challenges to the Initiative and long-term management challenges for DØ and the Computing Division. In the short-term, volunteer effort cannot be expected with certainty to follow a schedule required to coordinate contributions from many parties into a coherent release. It also exposes the goals of the Initiative to passive resistance if a disagreeing volunteer chooses to miss scheduled deliveries. In the long-term, volunteer efforts mask the need to identify long-term support for a mission-critical tool through an experiment-institution MOU. This exposes the experiment to the risk of not being able to support the application through the inevitable platform change (a major OS upgrade, for

example) or personnel change-over. This exposes CD to the risk of being asked to support the tool on short notice in order to continue to provide a stable, turnkey service for the experiment. Dealing with this sooner will simplify preparation for the long-term production-oriented maintenance and support envisioned by the RUN2 Initiatives.

Next Steps

While the Lessons Learned section covers a broad set of topics, long-term and short-term, the next steps that we believe may have the greatest impact are:

- A Service Level Agreement should be developed between CD and DØ which includes a maximum supported capacity based on existing experience and system configuration, with some cost associated with increases to this capacity (described in Lessons Learned/Production Operations Support)
- Monitoring should be identified or developed to support the SLA by tracking the parameters in the SLA to provide a history over time as well as fair warning in the instant if limits are approached or exceeded.
- Improved monitoring to support users with a single consistent view of the composite Grid Production System, such as a single job state diagram.
- Continue to identify higher priority development to support operations. Topics such as the “32k sub-directory limit and Grid tool behaviors which trigger this” will continue to require some developer work at a low level in order to reduce operations support load. Identifying problem frequencies in the ticket tracking system will guide cost-benefit prioritization of such work.
- Maintain the discipline of the new change request process. Continue the operator-developer-manager collaboration to clarify and enhance such processes as needed to manage the evolution of the system in a production context in the long-term.
- Continue the dialogue with RUN2 experiments on how to jointly achieve production-oriented operations which require reduced effort and less reliance on a small number of very experienced imbedded experts.
- Continue to pursue the installation and testing of LCG forwarding nodes in the standard configuration at two or more separate sites for redundancy.

Summary

With only limited additional effort to help coordinate existing groups, the DØ Grid Production Computing Initiative accomplished many of its goals, laying a solid foundation for future work. The production system was thoroughly retrofitted with new hardware and upgraded middleware, with a positive impact on the stability of the service (as of February 2008 at expected demand levels). Support for this “developer operated” system was transferred in a managed fashion to an already burdened operations group, with sufficient automation of routine tasks and documentation to make this feasible. Knowledge transfer activities and a new support process have been embraced by both operators and developers. Documentation and procedures at all levels have been created and updated to enable stable operations in the long-term. With much input from remote and local developers, a means to adapt DØ applications to the Grid Production has been deployed and documented which gives DØ more options for processing (with acceptable adaptation costs) and breaks reliance on the CAB system for several processing efforts. Generic WBS’s have been developed to help CD and DØ plan for this work in the future. While not all that was originally envisioned was accomplished (or accomplishable), the basic foundation of preparing DØ for reduced operations as Run2 winds down has been prepared as best as possible in the time allotted.